Parallel Processing of Large Datasets from NanoLC-FTICR-MS Measurements

Y. E. M. van der Burgt, I. M. Taban, M. Konijnenburg, M. Biskup, M. C. Duursma, R. M. A. Heeren, and A. Römpp*

FOM Institute for Atomic and Molecular Physics (AMOLF), Amsterdam, The Netherlands

R. V. van Nieuwpoort and H. E. Bal

Department of Computer Science, Free University, Amsterdam, The Netherlands

A new approach for automatic parallel processing of large mass spectral datasets in a distributed computing environment is demonstrated to significantly decrease the total processing time. The implementation of this novel approach is described and evaluated for large nanoLC-FTICR-MS datasets. The speed benefits are determined by the network speed and file transfer protocols only and allow almost real-time analysis of complex data (e.g., a 3-gigabyte raw dataset is fully processed within 5 min). Key advantages of this approach are not limited to the improved analysis speed, but also include the improved flexibility, reproducibility, and the possibility to share and reuse the pre- and postprocessing strategies. The storage of all raw data combined with the massively parallel processing approach described here allows the scientist to reprocess data with a different set of parameters (e.g., apodization, calibration, noise reduction), as is recommended by the proteomics community. This approach of parallel processing was developed in the Virtual Laboratory for e-Science (VL-*e*), a science portal that aims at allowing access to users outside the computer research community. As such, this strategy can be applied to all types of serially acquired large mass spectral datasets such as LC-MS, LC-MS/MS, and high-resolution imaging MS results. (J Am Soc Mass Spectrom 2007, 18, 152–161) © 2007 American Society for Mass Spectrometry

Nowadays mass spectrometry (MS) is the method of choice for the systematic analysis of a proteome [1]. Moreover, mass spectrometry–based proteomics is now one of the key players in systems biology, i.e., the integrated approach of different technical disciplines to study the physiological processes in a cell or tissue [2]. The number of researchers using MS for protein and peptide analyses is still rapidly increasing. Multiple instrumental developments have made MS accessible to a broader research community and enabled automatic data acquisition. In clinical research mass spectrometry has opened new ways of (early) detection of diagnostic biomarker molecules. Their identification is done by differential analysis of protein-expression patterns in patient and control samples. These patterns often change dramatically as a result of a disease and are thus helpful in early detection. Additionally, detection of such biomarkers can also play a significant role in prevention. In search for these biomarkers Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS) is a powerful tool because of its distinguishing feature of ultrahigh mass resolution. In a proteomics setup, FTICR-MS provides high mass precision and high mass accuracy of complex peptide mixtures, thus enabling peptide and protein identifications with high confidence. It is well known that the variation in protein concentration by more than ten orders of magnitude is one of the major challenges in proteomics [3]. The peptides that originate from high abundant proteins usually cause suppression of the low abundant peptide ions in electrospray ionization (ESI). A protein or peptide separation step such as gel electrophoresis or on-line liquid chromatography (LC) is necessary to reduce the complexity of the mixture. The FTICR mass spectrometer is perfectly suited for on-line coupling to a nanoLC-system provided adequate differential pumping is applied. NanoLC-FTICR-MS runs typically take 30–60 min and, with an ion cyclotron resonance (ICR) scan time of 1–2 s, this results in 900–3600 individual transients. These transients occupy 4 megabytes of disk space each and Fourier transformation is required to obtain corresponding mass spectra. The processing of these data has become a more time-consuming task than the LC-MS experiment itself and usually is carried out after the measurement. From these considerations the need for

Published online October 19, 2006

Address reprint requests to Prof. Ron M. A. Heeren, FOM Institute for Atomic and Molecular Physics, Macromolecular Ion Physics and Biomolecular Mass Spectrometry, Kruislaan 407, 1098 SJ Amsterdam, The Netherlands. E-mail: heeren@amolf.nl

* Present address: Justus Liebig University of Giessen, Institute for Inorganic and Analytical Chemistry, Schuberstr. 60, Bldg. 16, D-35392, Germany.

1044-0305/07/$32.00
doi:10.1016/j.jasms.2006.09.005

Received June 29, 2006
Revised September 6, 2006
Accepted September 14, 2006

automatic processing and improvement of speed is evident. In this paper we present new methodologies for processing large mass spectral datasets in a fully automated way. It will be shown that the total analysis time decreases dramatically upon using a flexible distributed computing environment. The increased processing speed enables on-line data analysis.

Parallel Processing of Large Mass Spectral Datasets

The analysis and interpretation of complex mass spectra has always been a challenging task for scientists in the field. Despite all modern computer facilities expert manual examination of a mass spectrum is still very common and necessary in an MS laboratory and thus remains an indispensable skill. However, it is evident that the amount and size of mass spectral datasets generated by modern mass spectrometers are incompatible with manual analysis. Examples of such large MS-based datasets are found in high-throughput proteomics experiments [4] as well as in high-resolution imaging MS experiments [5]. In a research lab real-time analysis of the experiments is pursued so that results can be used to adjust the parameters of the following experiment. Manual analysis of complex datasets is often inconsistent, incomplete, and error prone. Thus, the automation of mass spectral data analysis and interpretation of peptide profiling measurements is pivotal for the extraction of valuable information from each experiment and remains a key challenge in bioinformatics. Automation not only tackles the increasing data volumes but also allows repeated use of data analysis strategies with different parameters or datasets. This automated approach improves both flexibility and repeatability of analysis of large MS based datasets.

The need for automation was already recognized after the first ESI experiments [6] and has been further developed since [7–10]. Modules for distributing the computational MS/MS data searches have been described [11]. Here we present a new approach that combines preprocessing and postprocessing of serially acquired mass spectral datasets (e.g., LC-FTMS datasets) in a distributed computing environment. The speed of processing increases by making use of multiple connected computers instead of one. This type of processing is further referred to as *parallel* processing, resulting in a decreased total analysis time. The off-line processing of one single FT mass spectrum (in computer science referred to as a *job*) easily takes 3–4 s, mainly determined by the data transfer time and the peak picking algorithms. As a result, sequential processing of all jobs from one LC-FTMS dataset (such as 2000 spectra) amounts to a total processing time of at least 2 h. A parallel distribution of this workload (i.e., jobs) over different computers significantly decreases the total processing time. The requirements and details of a parallel setup are described in the experimental section. In short, a computer network (cluster) usually consists of machines (*nodes*) that run the same operating system and share a data storage facility. The server starts processing the raw data using the available nodes. In the final postprocessing step, the server summarizes the results and sends these to the data storage system as processed data, also referred to as *metadata*. In our definition, the metadata describes the original raw data on a higher abstraction level (i.e., processing results) in addition to how, when, and by whom the dataset was collected. Thus, all acquisition and processing parameters are stored, enabling tracking and reuse of all such variables. It is also possible to use different computer clusters simultaneously, which are managed by a central server. In this type of data processing a so-called grid approach is used. In a grid environment the nodes are platform independent and may be located at different geographical sites [12]. It is beyond the scope of this paper to discuss the intricate details of data processing using a grid. Here, the processing speed of large mass spectral datasets will be evaluated on both single processors (such as a desktop PC) and dedicated computer clusters. The work on automated data processing of large LC-FTMS datasets described in this paper was embedded in the Virtual Laboratory for e-science (VL-*e*). VL-*e* provides a science portal for distributed analysis, such as creation and submission of jobs on a distributed computer system in a grid. Furthermore, VL-*e* aims at allowing access to users outside the computer research community, thus facilitating new scientific collaborations in grid environments.

An alternative way to speed up the processing of large datasets is reducing the data during the measurement ("on-the-fly"). This approach is implemented in the hardware of the LTQFT. Here, the original measurement (raw data) is discarded and only the reduced mass spectra are saved. Clearly the advantage is that processing of the data is finished immediately after the measurement, thus reducing the total analysis time. Unfortunately, this step is not a loss-less procedure and excludes the possibility of reprocessing the raw data with a different set of parameters. Thus, the storage of all raw data is recommended, enabling future reanalysis or reprocessing with a new set of parameters (e.g., calibration, apodization) [13]. Here we demonstrate an approach that achieves full analysis on approximately the same timescale as achieved with on-the-fly processing but has the added advantage of safely storing the original raw data for future reprocessing.

As an example, the parallel data processing approach described was tested using an algorithm that was developed specifically for processing datasets obtained from nanoLC-FTICR experiments. The serial mass spectra from LC-FTICR-MS are perfectly suited for parallel processing because the spectra are (at least for the initial analysis) independent of each other—i.e., they can be analyzed separately. The setup of the algorithm is modular, which enables easy addition of

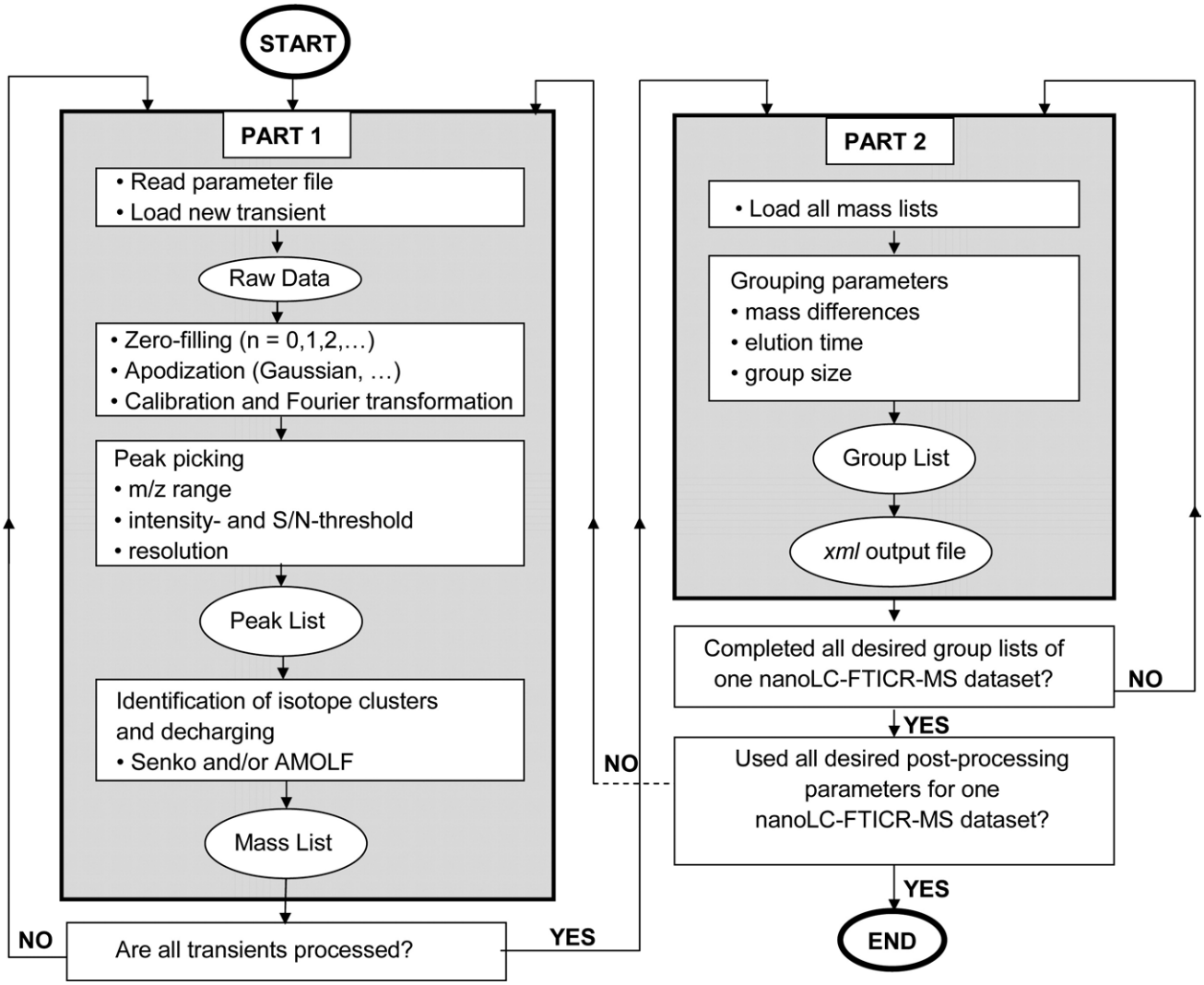

Figure 1. Structure and workflow of the PP-VLAM algorithms for parallel processing of datasets obtained from nanoLC-FTICR-MS experiments.

new processing modules or change of parameters or routines. In this way the algorithm can be used for all types of LC-MS and LC-MS/MS datasets. The modular (workflow) nature of this approach also enables automated processing of other types of large mass spectral datasets such as the results obtained from high-resolution mass spectral imaging experiments [5].

Methodology

Algorithm for Processing Large FTICR-MS Datasets

The main objective of this work is to enable processing of complex and large mass spectral datasets in a fully automated way using computational resources in the grid. To this end, an algorithm is developed that facilitates processing of large MS datasets in parallel. Basically, this PP-VLAM algorithm (Parallel Processing Virtual Laboratory AMsterdam) consists of two parts. In the first part a mass list for each mass spectrum is generated. The subsequent second part generates a summary for the full dataset (written as an *xml* output file). The contents of the first part of this algorithm depend on the type of mass spectrometer used, whereas the latter part of the algorithm is generic and thus applies to all kinds of different LC-MS datasets (e.g., obtained from quadrupole, ion trap, or time-of-flight instruments). Note that this approach also enables parallel processing of other two-dimensional mass spectral datasets, such as linescans in mass spectral images [5]. As an example, the PP-VLAM algorithm is described for datasets obtained from nanoLC-FTICR-MS experiments in Figure 1. In the first part a transient file (raw data) is located and, if necessary, transferred by the network to the actual processing computer. Then the transient is Fourier-transformed and calibrated, yielding a mass spectrum. The required processing parameters such as zero-filling, apodization, calibration, and

peak picking parameters have initially been user-defined. After the mass spectrum is obtained, different sets of isotopic peaks from a single molecular or fragment ion entity (i.e., isotopic cluster) are identified using either the Senko [7, 8] or our in-house–developed AMOLF routine (as described later). The monoisotopic mass of each identified isotopic cluster is calculated and summarized in a so-called mass list. These steps in the first part of the algorithm are repeated until all transients are processed and all mass lists have been completed. This approach and its corresponding algorithms is extremely suitable for massively parallel processing. The server collects all the data after processing all transients in parallel and uses a second postprocessing algorithm to group all mass lists to generate a so-called group list. In the following sections the design and the performance of the PP-VLAM algorithm are discussed in more detail.

Parameter File and File Transfer Protocols

The parameters for data processing are specified in a so-called properties file. In this file the location of the data is defined together with all user-defined parameters that are further used in the PP-VLAM algorithm (e.g., apodization, threshold, mass tolerance). Different file transfer protocols were implemented, such as the *windows network protocol*, *scp*, *sftp*, *gridftp*, or *mpicopy*, or by using the *grid application toolkit* (GAT, www.gridlab.org). The communication between the server and the compute nodes is by *remote method invocation* (rmi), implemented either in Java or the Ibis language.

Peak Picking, Noise Filtering, and the Peak List

Peaks are identified by searching through the mass spectrum until a value is found that exceeds a user-defined threshold. The area in the proximity of the *m/z* value where this occurs is subsequently searched for a local maximum (peak picking). The search area for this local maximum is defined by a mass width and the number of neighboring data points to verify. After the local maximum is found, the peak position (*m/z*) is defined at the middle of the full peak width at half height (FWHM). When the resolution of a peak is higher than the theoretical Fourier resolution ($(f*T_{transient})/2$ [14] the peak is considered as a noise signal and discarded. Note that all peaks that are not recognized as a local maximum are not further analyzed in the algorithm. It is therefore crucial to detect as many local maxima (peaks) as possible during the peak picking routine. The result of this approach is an extensive peak list.

Single Processing Results: Isotopic Cluster Identification, Decharging, and the Mass List

The peak list contains a significant amount of redundant information. A single peptide will generate different isotopomers and can present itself in different charge states. To reduce this complexity the peak list is converted into a mass list where all redundant information is reduced to one single mass entry for each compound. To achieve this a sequence of deisotoping and decharging modules is introduced.

Decharging of the isotopic clusters after determination of their charge is performed using either the Senko [7] or the Zscore algorithm [9, 10]. For the analysis of our peptide spectra, an in-house (AMOLF) developed routine was used for cluster identification and compared with the results from Senko and Zscore. The AMOLF routine matches a specific *m/z* distance $\Delta_{m/z}$ between two different peaks to a certain charge state q (q = 1, 2, 3···), defined as

$$\Delta_{m/z} = \left(M_{neutron}/q\right) * \left(1 \pm \sigma\right)$$

where $M_{neutron}$ is the mass difference (in Daltons) between two consecutive $^{12}C/^{13}C$ isotopic peaks belonging to the same compound, q is the charge state of the cluster, and σ is the user-specified precision with which this distance is used.

Using different values for σ enables the analysis of mass spectra obtained from different instruments with different peak resolutions. In addition, the ratio of the intensities of the first and second isotopes is determined, and should be in agreement with expected numbers of the compounds analyzed (e.g., for tryptic peptides this ratio is between approximately 3 and 0.3 for peptides with masses of, respectively, 500 and 5000 Da). The results from the two different decharging routines will be discussed in more detail in the results and discussion section.

After processing a certain transient (or mass spectrum) a list of masses is generated. All peaks from a spectrum are summarized in a table with their original charge state, their intensity and resolution (FWHM), and the position in the isotopic cluster. Note that at this stage a peptide (or any type of compound) that is detected with two (or more) different charge states in one mass spectrum results in two (or more) *almost* identical peptide masses in the mass list. The user defines whether these different mass determinations from the same peptide are averaged. When using sub-ppm mass accuracies the consideration of two different peptide masses of the *same* peptide proved to be important, as will be exemplified in the results and discussion section.

Final Processing Result: Grouping and the Group List

The first part of the algorithm (either sequential or parallel processing) results in a separate mass list for each spectrum. Depending on the type of experiment multiple mass lists may contain redundant information. These entries in the mass lists are grouped together to further reduce the complexity and redundancy of the

parallel processing results. In this second part of the process cycle two grouping criteria are used. A scan number range is defined to ensure that the masses found belong to the same LC-peak. A mass range is defined (or tolerance) to ensure that the mass spectral peaks found belong to the same molecule. The user may additionally define a minimum amount of scans in which a certain mass was detected ("group size"). In the case of chromatographic separation before mass analysis the mass of each eluting peptide is measured in different (sequential) scans if the LC peak is wide enough. Thus, different peptide masses are considered as a single eluting LC peak provided they are within a specific mass *and* time range. Moreover, this algorithm enables the selection of (consecutive) scans within an eluting LC peak and it is possible to generate a group list of a certain part from an LC run. The resulting group list, which is actually a peak list of all eluting peptides during one specific LC-FTICR experiment, can be used for further data analysis such as database searching.

xml Output File

The *xml* format is used for storage of the processed data. This format is encouraged by the HUPO Proteomics Standards Initiative (PSI) that defines community standards for data representation in proteomics [15, 16]. All metadata are stored in an *xml file*, i.e., all parameters that were used for processing the mass spectral dataset are documented. This *xml* file can be easily converted to the *mzdata* format [15]. Additionally, all the processing results from both the first and the second parts of the algorithm (mass lists and the group list) are stored in the same *xml* file. Each time a specific dataset is processed using different parameters a new *xml* file is generated, allowing for proper comparison between different processing results.

Experimental

NanoLC-FTICR-MS

The nanoLC system (LCPackings, Amsterdam, The Netherlands) consists of an autosampler, a switching unit, a nanoflow system, and UV detector. The switching unit is equipped with a reverse-phase capillary precolumn (C_{18} PepMap 100, internal diameter 0.3 mm, length 1 mm) and is used for preconcentration of the sample at a flow rate of 30 μL/min. Peptide separation is then carried out on an analytical column (PepMap 100, internal diameter 0.075 mm, length 15 cm) using nanoflow elution at 300 nL/min. Typically, the injection volume was 2 μL. The eluents used were 1% acetic acid and 5% acetonitrile in water (A) and 1% acetic acid and 10% water in acetonitrile (B). The gradient used for the separation of peptides was: 0–30 min, 0–50% B, followed by 30–35 min, 50–90% B. The nanospray source connecting the LC system to the mass spectrometer was built in-house and equipped with New Objective Picotips™.

All the peptide mass measurements were performed in the positive ion mode using a modified APEX 7.0eT FTICR-MS (Bruker Instruments, Billerica, MA), equipped with a 7 T superconducting magnet and an infinity cell [17]. The ions generated by the electrospray ion source are accumulated in an octopole ion-trap (typical accumulation time 0.4 s) before being transferred to the ICR cell by two quadrupole ion guides. The ions were trapped in the ICR cell using side-kick. In this way, a typical scan time was 1.3 s. All experimental parameters were controlled using software and hardware developed in-house as part of the continual evolution of this proteomics/fundamental studies instrument.

The results of the nanoLC-FTICR-MS experiments are displayed in the AWE3D module (Arbitrary Waveform Editor), which is part of the AWTools software package [17]. This software is written in C++ and can be used for a first evaluation of chromatographic separation, mass spectral resolution, and sensitivity/intensity. It displays total and selected ion currents as well as a three-dimensional representation of the data. The whole dataset can be recalibrated or apodized, zero-filled in this mode.

Samples and Protein Identifications

NanoLC-FTICR-MS measurements of two different protein samples were used to verify the performance and outcome of the PP-VLAM algorithm: (1) a tryptic digest of 20 μM/mL savinase (Sigma, St. Louis, MO) and (2) a tryptic digest of a protein mixture (50 μM/mL BSA (Sigma), 50 μM/mL ovalbumin (Sigma), and 50 μM/mL lysozyme (Fluka Chemie GmbH, Deisenhofen, Germany). For protein identification, the peptide masses were submitted to a Mascot database search (MatrixScience, London, UK) with a mass tolerance of 20 ppm using the SwissProt database.

Also, different clinical cerebrospinal fluid (CSF) samples from breast cancer patients with leptomeningeal metastasis (brain tumor) were used. The control samples originated from headache patients without brain tumors. All samples were subjected to trypsin digestion (Promega, Madison, WI) after addition of 0.2% Rapigest (Waters Associates, Milford, MA) in a 50 mM ammoniumbicarbonate buffer. These CSF samples were provided by the Erasmus Medical Center (EMC) in Rotterdam and have been part of a more extensive biomarker study based on MALDI-TOF measurements [18].

Programming Software PP-VLAM and Parallel Processing

The PP-VLAM algorithm is written in Java and all features are Java implementations, i.e., the algorithm is platform independent and thus runs on every operating system. In principle, this enables the use of all comput-

ers that are not used at full capacity in the laboratory. The PP-VLAM algorithm will be made available to the scientific community through the Internet within the framework of the Virtual Laboratory for E-science (VL-*e*). Because of its modular setup this software is not limited to the FTICR data discussed herein but can be easily adjusted for a variety of mass spectral datasets.

For parallel processing of mass spectral datasets several workers (computer nodes) request jobs (each consisting of one spectrum to be processed) upon their availability. All jobs are managed by a central server and the computational resources are linked by TCP/IP connections. The Java program code is stored centrally on a file server to ensure that every computer node will access the same version. The properties file (containing the user-specified set of processing parameters) can be stored in any location accessible by the server. The processing details are only read by the server module, which passes them on to the worker modules. Each node transfers and processes raw data (mass spectral file) separately and returns the results to the server. The server and worker modules can run either on one or on separate computers. The server and worker programs can be started manually, in batch or by any other system that has the ability to execute programs, such as a grid job. Each worker requests the processing parameters as specified in the properties file once. On receipt of a job request the server responds with details about a job, such as the filename. Then the worker loads the corresponding mass spectrum from the data storage system. When the worker has processed the job, it returns the resulting mass list to the server and requests a new job. The server stops and compiles a concise report when all jobs have been processed. The final result list is generated as an *xml* file (as described earlier) that is stored locally or on the central data storage system.

Computer Hardware

The measurement data are stored on a 5.5 TB storage system (SGI Origin 300, Lexington, MA) that is linked by a 2 Gigabit fiber-channel connection to the acquisition computer (FTICR-MS acquisition). Different computer platforms were used to test the performance of the analysis software. The single desktop PC was a Pentium4 processor (3.2 GHz) with Hyper Threading and 1 GB RAM. The PC cluster consisted of five personal computers running either Windows NT, 2000, or XP, or Linux as an operating system. The internal computer cluster that was used is located at AMOLF and consists of 38 compute nodes, each equipped with dual AMD opteron 2.2 GHz processors. The nodes are internally linked by 1 gigabit Ethernet connections and externally by a 1 gigabit glass-fiber connection. The Dutch national computer cluster that was used ("Lisa") is located at SARA (Dutch Supercomputing Center [19]) and consists of 630 compute nodes, each equipped with dual Xeon 3.4 GHz processors. The nodes are linked by 1 gigabit Ethernet connections and run on a Debia Linux operating system.

Results and Discussion

Evaluation of the PP-VLAM Algorithm Using Protein Standards

Two different nanoLC-FTICR mass spectral datasets were used to evaluate the performance of the described PP-VLAM algorithm. Five peptides from the resulting group list from a tryptic digest of savinase were assigned within a mass accuracy of 3 ppm after internal calibration. The total sequence coverage (s.c.) in this case is 30%. From a tryptic digest of a protein mixture (containing three well-defined proteins; see experimental section), 17 peptides from BSA (s.c. = 31%), 11 peptides from ovalbumin (s.c. = 41%), and 6 peptides from lysozyme (s.c. = 56%) from the group list were assigned all within a mass accuracy of 20 ppm. The lower mass accuracy of peptides in the protein mixture compared to savinase partly results from peptide concentration differences in the protein mixture. For example, peptides that elute in high concentration cause overloading of ions in the ICR cell and thus increased mass shifts. Moreover, in the protein mixture the relative intensity of a single peptide compared to the total ion intensity at different time points varies between 5% and almost 100% negatively affecting the mass measurement accuracy.

In conclusion, these results demonstrate the ability of the PP-VLAM algorithm to process complex nanoLC-FTICR data in an efficient way. The application of the PP-VLAM algorithm to a peptide mixture generates a peak list that is very well suited for database searches and thus protein identification.

Evaluation of the Senko and AMOLF Routines for Isotope Cluster Identification

For cluster identification and decharging of peptide peaks the Senko and AMOLF routines both proved to be extremely powerful (in terms of computing speed and total amount of identified clusters) compared to the Zscore- and averaging-based routine [9]. The latter two methods are more suited for mass spectral analysis of intact proteins. A detailed comparison between the Senko and AMOLF routines was performed using the standard protein datasets described earlier. For savinase, all five identified tryptic peptides were found using either the Senko or AMOLF routine for isotope cluster identification. For the protein mixture, all 34 identified tryptic peptides were found with both the Senko and AMOLF routines. Additionally, using the AMOLF routine two more peptides (from BSA) were identified. In general, the mass list contained roughly 20% more peptides using the AMOLF routine compared to the Senko algorithm. This results from the less-stringent requirement that only two peaks above a defined threshold are sufficient for an isotope cluster in the AMOLF routine, whereas in the Senko routine all other peaks in the cluster area are also taken into account.

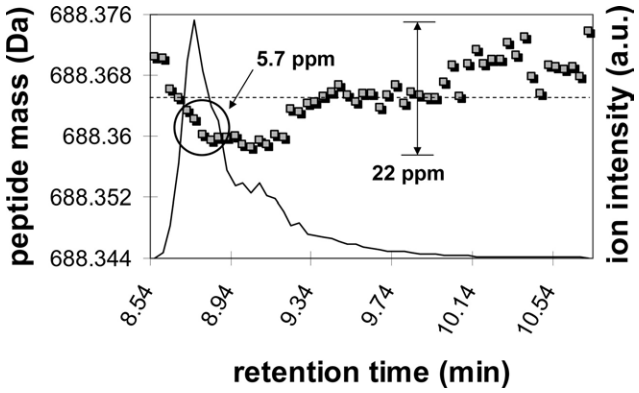

Figure 2. NanoLC-elution profile of a singly protonated BSA peptide. The dotted line indicates the theoretical mass of this peptide, the squares show the FTICR-measured masses in each scan. Upon post-acquisition selection of scan numbers the difference between the highest and lowest measured peptide mass decreases 4-fold, thus improving scan precision.

Scan Precision and Mass Accuracy

Usually, each eluting chromatographic peak is mass-analyzed multiple times during an LC-MS experiment. In practice, each mass spectrum (scan) from one peptide results in a slightly different peptide mass. This variation in detected peptide mass, further referred to as *scan precision*, is dependent on different factors such as the type of instrument used, the chromatographic resolution, and the ion intensities itself. As mentioned earlier, the high mass accuracy of FTICR-MS is a well-distinguished feature. In modern FTICR analyzers this accuracy has improved toward sub-ppm levels. Clearly, high mass accuracy improves the reliability of database searches and, moreover, it helps to identify peptides that are not in a database (e.g., unsequenced species, post-translational modifications) [20, 21]. The advantage of the PP-VLAM algorithm compared to other processing software packages is that all the different peptide masses are stored in the *xml* file and thus can be analyzed in more detail. In this way the precision of the data can be significantly improved after data acquisition. As an example, in Figure 2 the elution profile of a BSA peptide (theoretical mass 688.3656 Da) in the protein mixture is shown. The peptide masses (shown as squares) are measured after external calibration. The difference between the highest and lowest measured mass is 22 ppm. Clearly, in the tail of the eluting peptide the variation increases. Upon user-defined selection of scan numbers (retention time), as indicated with the circle, the scan precision improves fourfold. This information can be further used in an iterative way to internally recalibrate the spectra and thus improve the mass accuracy after the data acquisition. Often a peptide is mass analyzed at two (or even more) different charge states. From Figure 3 it can be seen that using the PP-VLAM algorithm a singly and a doubly protonated BSA peptide can be analyzed separately. In this case, the scan precision of the doubly protonated species can be improved to 3.6 ppm.

Visualization of Peptide Profiles from CSF Samples: "A Real-Life Example"

More advantages of the processing and analysis tools described herein are exemplified using a dataset obtained from nanoLC-FTICR-MS measurements of cerebrospinal fluid (CSF) samples. The cerebrospinal fluid encloses the brain and is thus an ideal medium to investigate diseases that affect the central nervous system (CNS) such as Alzheimer's disease or brain tumors. Details of the samples are given in the experimental section. In general, the objective is to compare peptide profiles of healthy and diseased individuals for detection of possible biomarkers. Each nanoLC-FTICR-MS measurement yields a large and complex dataset, and processing and visualization of each dataset are pivotal for proper comparison and thus extraction of valuable information.

As an example, all sequential FTICR mass spectra acquired during one nanoLC-experiment of a CSF sample are shown in Figure 4a. Obviously, the manual comparison of such plots is extremely tedious and error-prone. Application of the PP-VLAM algorithm to this specific dataset results in an *xml* output file that contains between 100 and 200 peptide masses (depending on the processing parameters). Clearly, such a list of peptide masses can be easily compared with those derived from replicate measurements (repeatability of the LC-MS measurement of one sample) or with other samples (e.g., patient and control comparison). For comparison of the *xml* output files we developed a tool for visualization of either the mass lists or group lists from one or multiple samples (see Figure 1 for explanation of the terms mass and group list). In Figure 4b an example is shown for a CSF sample (same as in Figure 4a) processed with different parameters. In this case the specific dataset is visualized for apodized and nonapodized raw data, clearly showing the similarities and differences between these two types of processing. The complexity of this plot is far less compared to the spectral data in Figure 4a. In addition, vertical lines that

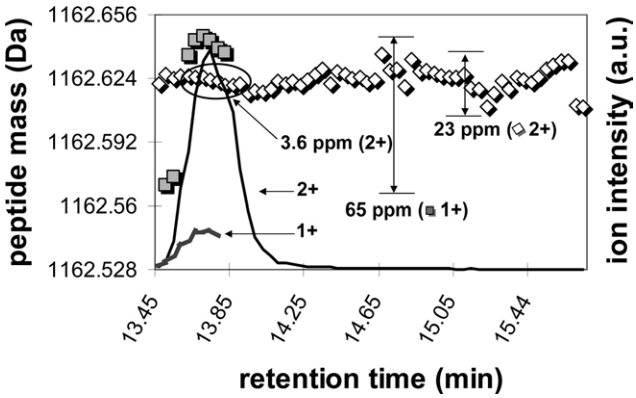

Figure 3. NanoLC-elution profile of a singly and doubly protonated BSA peptide. The different charge states of the *same peptide* result in different scan precisions. It is thus recommended to analyze these *different species* separately.

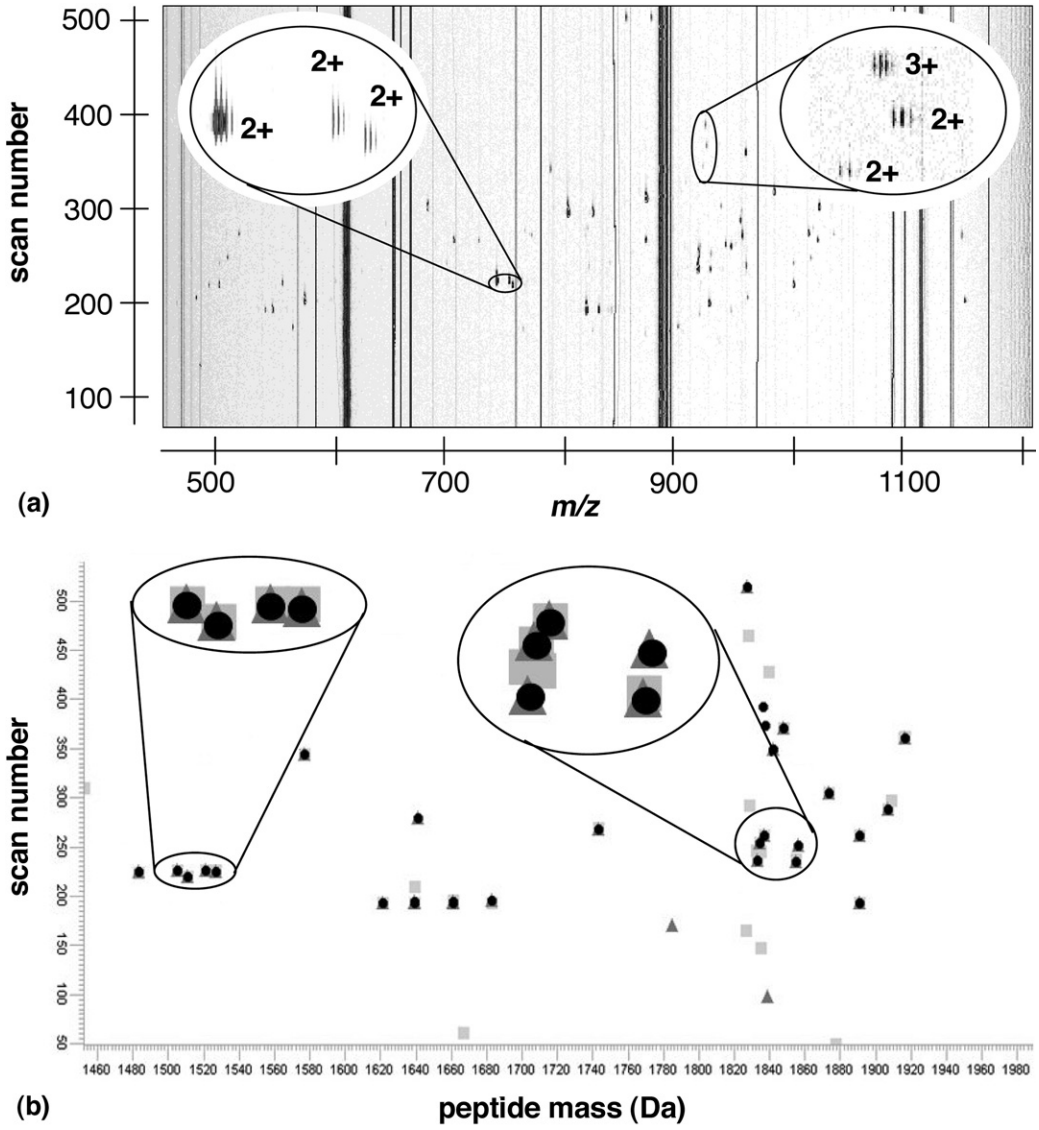

Figure 4. (**a**) NanoLC-FTICR-MS measurement of a trypsin digested cerebral spinal fluid (CSF) sample displayed with the AWE3D software. The *m/z*-values are plotted on the *x*-axis, the scan numbers (linear with nanoLC retention times) on the *y*-axis and the peak intensities on the *z*-axis (out-of-plane). (**b**) Visualization of a processed nanoLC-FTICR-MS measurement of a digested cerebral spinal fluid (CSF) sample. The peptide masses are plotted on the *x*-axis (from 1450 to 1980 Da) and the scan numbers (corresponding to retention time) on the *y*-axis (50-450). A peptide mass is visualized provided that it is detected in at least 3 sequential scans with a mass precision σ of 0.05 Da (see methodology). Here, different processing results are overlayed in one plot, *i.e.* no apodization on the raw data (light grey squares), apodization and AMOLF isotope cluster identification (grey triangles) and apodization and Senko isotope cluster identification (black circles). The inserts show the similarity between AMOLF and Senko cluster identification results (*i.e.* each triangle is overlapped by a circle). However, without apodization additional clusters are detected or the cluster is observed at a slightly different peptide mass.

result from chemical and electronic noise in Figure 4a are efficiently removed in the processed dataset. In a similar way, this viewer enables visual comparison of replicate measurements or of different samples.

Speed Benefits of Distributed Computing of Large Mass Spectral Datasets

Several hardware configurations ranging from a single desktop computer to supercomputers (clusters) were evaluated for parallel processing of large mass spectral datasets. For a dataset obtained from nanoLC-FTICR-MS of the protein mixture (877 spectra, 3.4 gigabyte) the total processing time was more than 2 h using a single desktop computer. Using a cluster of five desktop computers the processing time decreases to 0.5 h, although this time is limited by the file transfer speed (i.e., the speed depends on the quality of the internal network between the computers and the central

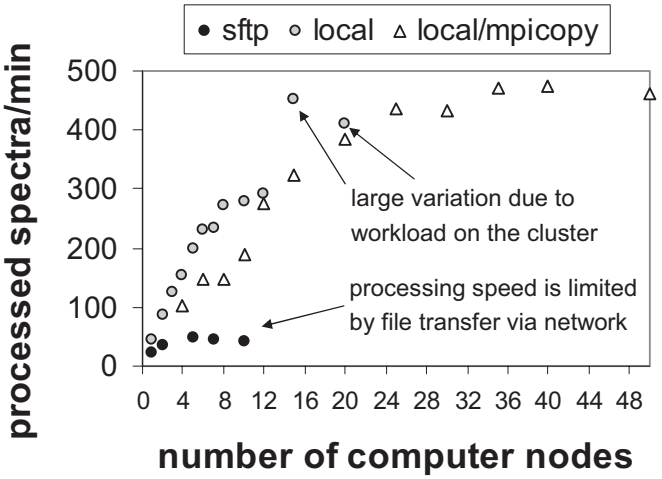

Figure 5. Increase in processing speed of a large mass spectral dataset using a distributed computing environment. The speed in the cluster is limited by the *sftp* transfer speed of raw data from the central storage facility to the computer nodes (black circles). The processing speed is improved up to 12 computer nodes when the raw data is stored on the cluster itself (grey circles). Note that the total processing time depends on the other activities of the cluster when using more than 15 nodes. This variation in processing time is smaller when using an *mpicopy* protocol [19] where up to 50 computer nodes can be assigned (white triangles).

raw data storage). A summary of the results on different clusters is given in Figure 5. Clearly even on a dedicated computer cluster the amount of spectra processed per minute is limited by the network speed. This limitation in speed up of distributed computing is well known from computer sciences. The transfer of raw data from the central storage site outside the cluster to the multiple computer nodes becomes inefficient using more than seven nodes. As an alternative, the total raw dataset from an LC-MS experiment can be locally stored on the shared storage of the cluster itself or copied temporarily to all individual nodes using the *mpicopy* protocol (this takes less than 3 min [19]). In the latter case the total 3.4 gigabyte dataset is processed on 40 computer nodes within 2 min, resulting in a total processing time of less than 5 min.

Conclusions

Automatic processing of large mass spectral datasets in a distributed computing environment allows for a substantial reduction of the analysis time. We showed full processing of a 3 gigabyte raw dataset from a nanoLC-FTICR-MS experiment is within 5 min using an in-house–developed algorithm on a dedicated computer cluster. Because of its modular setup the algorithm can be applied to all other types of hyphenated or serial mass spectral datasets (e.g., LC-MS, LC-MS/MS, TOF imaging MS). In our approach the storage of all data is preferred to discarding raw data during the measurement, thus enabling future reanalysis or reprocessing using a new set of parameters. Furthermore, automatic processing improves the repeatability of the analysis over the more error-prone manual analysis. The ability to reuse the parallel processing modules described in this paper in a distributed workflow environment allows for new scientific collaborations to be realized in the virtual laboratory. This in turn enables the scientists to share processing tools and strategies and enhances the quality of experimentation with large mass spectral datasets.

Acknowledgments

This work was carried out within the framework of the Netherlands Proteomics Centre (NPC) and the Virtual Laboratory for E-science (VL-*e*). The work is part of the research program of the "Stichting voor Fundamenteel Onderzoek der Materie (FOM)," which is financially supported by the "Nederlandse organisatie voor Wetenschappelijk Onderzoek (NWO)." The authors thank Theo Luider (Erasmus Medical Center, Rotterdam, the Netherlands) for providing the CSF samples.

References

1. Aebersold, R.; Mann, M. Mass Spectrometry-Based Proteomics. *Nature* **2003,** *422,* 198–207.
2. Ideker, T. T.; Thorsson, V.; Ranish, J. A.; Christmas, R.; Buhler, J.; Eng, J. K.; Bumgarner, R.; Goodlet, D. R.; Aebersold, R.; Hood, L. Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. *Science* **2001,** *292,* 929–934.
3. Anderson, N. L.; Polanski, M.; Pieper, R.; Gatlin, T.; Tirumalai, R. S.; Conrads, T. P.; Veenstra, T. D.; Adkins, J. N.; Pounds, J. G.; Fagan, R.; Lobley, A. The Human Plasma Proteome: A Non-Redundant List Developed by Combination of Four Separate Sources. *Mol. Cell. Proteomics* **2004,** *3,* 311–326.
4. Masselon, C.; Pasa-Tolic, L.; Tolic, N.; Anderson, G. A.; Bogdanov, B.; Vilkov, A. N.; Shen, Y.; Zhao, R.; Qian, W. J.; Lipton, M. S.; Camp, D. G.; Smith, R. D. Targeted Comparative Proteomics by Liquid Chromatography–Tandem Fourier Ion Cyclotron Resonance Mass Spectrometry. *Anal. Chem.* **2005,** *77,* 400–406.
5. Heeren, R. M. A. Proteome Imaging: A Closer Look at Life's Organization. *Proteomics* **2005,** *5,* 4316–4326.
6. Mann, M.; Meng, C. K.; Fenn, J. B. Interpreting Mass Spectra of Multiply Charged Ions. *Anal. Chem.* **1989,** *61,* 1702–1708.
7. Senko, M. W.; Beu, S. C.; McLafferty, F. W. Automated Assignment of Charge States from Resolved Isotopic Peaks for Multiply-Charged Ions. *J. Am. Soc. Mass Spectrom.* **1995,** *6,* 52–56.
8. Senko, M. W.; Beu, S. C.; McLafferty, F. W. Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions. *J. Am. Soc. Mass Spectrom.* **1995,** *6,* 229–233.

9. Zhang, Z. Q.; Marshall, A. G. A Universal Algorithm for Fast and Automated Charge State Deconvolution of Electrospray Mass-to-Charge Ratio Spectra. *J. Am. Soc. Mass Spectrom.* **1998,** *9,* 225–233.
10. Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. Automated Reduction and Interpretation of High Resolution Electrospray Mass Spectra of Large Molecules. *J. Am. Soc. Mass Spectrom.* **2000,** *11,* 320–332.
11. Matthiesen, R.; Trelle, M. B.; Højrup, P.; Bunkenborg, J.; Jensen, O. N. VEMS. 3.0: Algorithms and Computational Tools for Tandem Mass Spectrometry Based Identification of Post-translational Modifications in Proteins. *J. Proteome Res.* **2005,** *4,* 2338–2347.
12. Kesselman, C.; Foster, I., Eds. *The Grid: Blueprint for a New Computing Infrastructure*; Morgan Kaufmann: San Francisco, CA, **1999**.
13. Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskii, A. The Need for Guidelines in Publication of Peptide and Protein Identification Data. *Mol. Cell. Proteomics* **2004,** *3,* 531–533.
14. Amster, I. J. Fourier Transform Mass Spectrometry. *J. Mass Spectrom.* **1996,** *31,* 1325–1337.
15. See the website: http://psidev.sourceforge.net/ms/index.html.
16. Orchard, S.; Hermjakob, H.; Apweiler, R. The Proteomics Standards Initiative. *Proteomics* **2003,** *3,* 1374–1376.
17. Taban, I. M.; McDonnell, L. A.; Römpp, A.; Cerjak, I.; Heeren, R. M. A. SIMION. Analysis of a High Performance Linear Accumulation Octopole with Enhanced Ejection Capabilities. *Int. J. Mass Spectrom.* **2005,** *244,* 135–143.
18. Dekker, L. J.; Boogerd, W.; Stockhammer, G.; Dalebout, J. C.; Siccama, I.; Zheng, P.; Bonfrer, J. M.; Verschuuren, J. J.; Jenster, G.; Verbeek, M. M.; Luider, T. M.; Smitt, P. A. MALDI–TOF Mass Spectrometry Analysis of Cerebrospinal Fluids Tryptic Peptide Profiles to Diagnose Leptomeningeal Metastases in Breast Cancer Patients. *Mol. Cell. Proteomics* **2005,** *4,* 1341–1349.
19. SARA Dutch Supercomputing Center, www.sara.nl. For details on the *mpicopy* protocol see www.sara.nl/userinfo/lisa/usage/progavail/index.html, "special utilities."
20. Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Pasa-Tolic, L.; Shen, Y.; Conrads, T. P.; Veenstra, T. D.; Udseth, H. R. An Accurate Mass Tag Strategy for Quantitative and High-Throughput Proteome Measurements. *Proteomics* **2002,** *2,* 513–523.
21. Spengler, B. De Novo Sequencing, Peptide Composition Analysis, and Composition-Based Sequencing: A New Strategy Employing Accurate Mass Determination by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2004,** *15,* 704–715.